Recent advances in natural language processing using neural networks have given rise to numerous methods of obtaining continuous-space vector representations of textual data that can be exploited for various applications. One of these methods is to use internal representations learned by neural machine translation (NMT) models. However, the attention mechanism in modern NMT systems removes the single point in the neural network from which the source sentence representation can be extracted. In this thesis, we propose and empirically evaluate novel ways to remove this limitation. We review existing methods of obtaining sentence representations and evaluating them, and present novel intrinsic evaluation metrics. Next, we describe our modifications to attention-based NMT architectures that allow extracting sentence representations. In the experimental section, we analyze these representations and evaluate them using a wide range of metrics with a focus on meaning representation. The results suggest that the better the translation quality, the worse the performance on these tasks. We also observe no performance gains from using multi-task training to control the representations.